

AD-A170 370

ESTIMATION FROM BINOMIAL DATA WITH CLASSIFIERS OF KNOWN  
AND UNKNOWN IMPER. (U) MARYLAND UNIV COLLEGE PARK DEPT  
OF MANAGEMENT SCIENCES AND ST. N L JOHNSON ET AL.

1/1

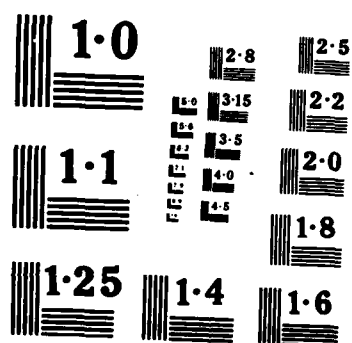
UNCLASSIFIED

JUL 86 UMD/HSS/1986/6 N00014-84-K-0301

F/G 12/1

NL





AD-A170 370

3

ESTIMATION FROM BINOMIAL DATA WITH CLASSIFIERS OF  
KNOWN AND UNKNOWN IMPERFECTIONS

Norman L. Johnson  
University of North Carolina  
Chapel Hill, North Carolina

and

Samuel Kotz  
University of Maryland  
College Park, Maryland

ABSTRACT: Observations from inspection by a 'test' method and a standard method are combined to provide estimators of population proportion, and of probabilities of misclassification for the test method. Results of Hochberg and Tenenbein [3] and of Albers and Veldman [1] are extended to the case where the standard method is not perfect, but its misclassification probabilities have known values. Both moment and maximum likelihood estimators are considered and some asymptotic properties of the resulting estimators are compared.

JULY 1986

DTIC  
ELECTE  
JUL 30 1986  
S D  
B

Key Words and Phrases: Binomial distribution; information matrix;  
inspection errors; maximum likelihood; method of moments; EM algorithm;  
statistical differentials.

DISTRIBUTION STATEMENT A

Approved for public release  
Distribution Unlimited

079

DTIC FILE COPY

## 1. INTRODUCTION

Suppose we have a large population, containing an unknown proportion,  $P$ , of individuals possessing a certain characteristic, which we will call 'nonconformance'. In a random sample, of size  $n$ , from this population, the distribution of the number,  $X$ , say, of nonconforming individuals will be binomial with parameters  $n, P$  so that

$$\Pr[X=x] = \binom{n}{x} P^x (1-P)^{n-x} \quad (x = 0, 1, \dots, n).$$

We will represent this, symbolically, as

$$X \sim \text{Bin}(n, P), \text{ where } \sim \text{denotes "is distributed as".}$$

If the individuals in the sample of size  $n$  are examined by an imperfect measuring device, which detects actual nonconformance with probability  $p$ , and (incorrectly) 'detects' nonconformance, when the individual is really not nonconforming, with probability  $p'$ , then the distribution of  $Z$ , the number of individuals declared to be nonconforming, as a result of this inspection, will be binomial with parameters  $n$ ,  $Pp + (1-P)p'$ . It is clear that the only parameter that can be estimated from observations on values of  $Z$  in independent samples is  $Pp + (1-P)p'$ .

Various methods have been suggested for obtaining data from which estimates of  $P$ ,  $p$  and  $p'$  can be derived (e.g. Albers and Veldman [1], Johnson and Kotz [5]). Tenenbein [6] suggested additional inspection of part of the sample by a perfect measuring device (for which  $p=1$  and  $p'=0$ ) and utilizing the resultant data. This method has been extended by Hochberg and Tenenbein [3] to allow for inspection of a further sample, of size  $n_S$ , say, by the perfect measuring device ( $S$ ).

In this paper, we study problems arising in this latter situation if the 'established' measuring device S is not perfect, but has known values  $p_S, p'_S$  for  $p, p'$  respectively. For convenience, we will denote the (unknown) values of  $p, p'$  for the measuring device under test (T) by  $p_T, p'_T$  respectively. We will also assume (when necessary) that  $p_S \gg p'_S$  and  $p_T \gg p'_T$ .

Problems of this kind arise when it is desired to calibrate the new device (T), by estimating  $p_T$  and  $p'_T$ . The unknown proportion (P) of NC units plays the role of a nuisance parameter in such problems.

## 2. ANALYSIS I (Moment Estimation)

As a consequence of the inspections we have the following sets of observations:

- (i)  $n_S$  using S alone, with  $Z_S$  judged nonconforming (NC),
- (ii)  $n_T$  using T alone, with  $Z_T$  judged NC,
- (iii)  $n$  using both S and T, with results shown below:

S \ T	# NC	# not NC
# NC	$Z_{11}$	$Z_{10}$
# not NC	$Z_{01}$	$Z_{00}$

(# denotes 'number of'.) Evidently,  $Z_{11} + Z_{10} + Z_{01} + Z_{00} = n$ .

Under the assumption of random sampling from a population of effectively infinite size, we have that:

$$Z_S, Z_T \text{ and } \tilde{Z} = \begin{pmatrix} Z_{11} & Z_{10} \\ Z_{01} & Z_{00} \end{pmatrix} \text{ are mutually independent;} \quad (1.1)$$

$$Z_S \sim \text{Bin}(n_S, \theta_S) \text{ with } \theta_S = p_S P + p'_S (1-P); \quad (1.2)$$

$$Z_T \sim \text{Bin}(n_T, \theta_T) \quad \text{with} \quad \theta_T = p_T P + p'_T(1-P), \quad (1.3)$$

Also, assuming that the S and T classifications are independent, given the true status of the individual,

$$Z \sim \text{Multinomial} \left[ n; \begin{pmatrix} \phi & \theta_S - \phi \\ \theta_T - \phi & 1 - \theta_S - \theta_T + \phi \end{pmatrix} \right] \quad (1.4)$$

with  $\phi = p_S p_T P + p'_S p'_T(1-P)$ , where  $\sim$  denotes "is distributed as".

Recall that  $p_S$  and  $p'_S$  have known values, and  $P$  is the (unknown) proportion of NC individuals in the population.

$$\text{Also} \quad P = (\theta_S - p'_S) / (p_S - p'_S) \quad (2.1)$$

$$p_T = (\phi - p'_S \theta_T) / (\theta_S - p'_S) \quad (2.2)$$

$$p'_T = (p_S \theta_S - \phi) / (p_S - \theta_S) \quad (2.3)$$

$$\text{Now,} \quad (n_S + n) \tilde{\theta}_S = Z_S + Z_{10} + Z_{11} \sim \text{Bin}(n_S + n, \theta_S) \quad (3.1)$$

$$(n_T + n) \tilde{\theta}_T = Z_T + Z_{01} + Z_{11} \sim \text{Bin}(n_T + n, \theta_T) \quad (3.2)$$

$$n \tilde{\phi} = Z_{11} \sim \text{Bin}(n, \phi) \quad (3.3)$$

so that  $\tilde{\theta}_S$ ,  $\tilde{\theta}_T$  and  $\tilde{\phi}$  (as defined in (3.1)-(3.3)) are unbiased estimators of  $\theta_S$ ,  $\theta_T$  and  $\phi$  respectively.

$$\text{Hence} \quad \tilde{P} = (p_S - p'_S)^{-1} (\tilde{\theta}_S - p'_S) \quad (4.1)$$

is an unbiased estimator of  $P$ . Although the estimators

$$\tilde{p}_T = (\tilde{\theta}_S - p'_S)^{-1} (\tilde{\phi} - p'_S \tilde{\theta}_T) \quad (4.2)$$

$$\text{and} \quad \tilde{p}'_T = (p_S - \tilde{\theta}_S)^{-1} (p_S \tilde{\theta}_T - \tilde{\phi}) \quad (4.3)$$

are not unbiased estimators of  $p_T$  and  $p'_T$  respectively, the biases should not be large if sample sizes are adequate (see the example later in this section).

The variance-covariance matrix of the random variables in (3.1)-(3.2) is

$$\text{Var}((n_S+n)\tilde{\theta}_S, (n_T+n)\tilde{\theta}_T, n\tilde{\phi}) = \begin{pmatrix} (n_S+n)\theta_S(1-\theta_S) & n(\phi-\theta_S\theta_T) & n\phi(1-\theta_S) \\ n(\phi-\theta_S\theta_T) & (n_T+n)\theta_T(1-\theta_T) & n\phi(1-\theta_T) \\ n\phi(1-\theta_S) & n\phi(1-\theta_T) & n\phi(1-\theta) \end{pmatrix}$$

Hence (cf. (4.1))

$$\text{var}(\tilde{P}) = (n_S+n)^{-1}(p_S-p'_S)^{-2}\theta_S(1-\theta_S) \quad (5.1)$$

and, using the method of statistical differentials (see, e.g. Johnson and Kotz [4, Chapter 1, Section 7.5]) we obtain, after some algebraic manipulation; the approximate formula

$$\begin{aligned} \text{var}(\tilde{p}_T) \doteq p_T^2 p^{-2} (p_S - p'_S)^{-2} & \left[ \left\{ n^{-1} \phi(1-\phi) - 2(n_T+n)^{-1} p'_S \phi(1-\theta_T) + (n_T+n)^{-1} p'^2_S \theta_T(1-\theta_T) \right\} p_T^{-2} \right. \\ & \left. - 2(n_S+n)^{-1} \{ \phi(1-\theta_S) - n(n_T+n)^{-1} p'_S (\phi - \theta_S \theta_T) \} p_T^{-1} + (n_S+n)^{-1} \theta_S(1-\theta_S) \right] \quad (5.2) \end{aligned}$$

An approximate expression for  $\text{var}(\tilde{p}'_T)$  is obtained from (5.2) by replacing  $p_T$  by  $p'_T$  and  $P$  by  $(1-P)$ , and interchanging  $p_S$  and  $p'_S$ .

An approximate formula for the bias of  $\tilde{p}'_T$  is

$$E[\tilde{p}'_T] - p'_T \doteq p'_T \left\{ \frac{\text{var}(\tilde{\theta}_S)}{(\theta_S - p'_S)^2} - \frac{\text{cov}(\tilde{\theta}_S, \tilde{\phi} - p'_S \tilde{\theta}_T)}{(\theta_S - p'_S)(\phi - p'_S \theta_T)} \right\} \quad (6)$$

which, after some reduction, gives a proportional bias (i.e.  $100(\text{bias})/p'_T\%$ )

$$\frac{100\{n_T p'_S(1-\theta_S) + n(1-p'_S)\theta_S\}(\phi - \theta_S \theta_T)}{(n_S+n)(n_T+n)(\theta_S - p'_S)^2(\phi - p'_S \theta_T)} \% \quad (7)$$

From (2.1)-(2.3)

$$\theta_S - p'_S = (p_S - p'_S)P; \quad \phi - p'_S \theta_T = p_T(\theta_S - p'_S)P$$

and also  $\phi - \theta_S \theta_T = P(1-P)(p_S - p'_S)(p_T - p'_T)$ , so

the approximate proportional bias (7) is

$$\frac{100\{n_T p'_S(1-\theta_S) + n(1-p'_S)\theta_S\} (1-P)(p_T - p'_T)}{(n_S+n)(n_T+n)P^2(p_S - p'_S)^2 p_T} \% \quad (7)'$$

which is positive and (since  $p'_T < p_T$ ) less than

$$\frac{100 G(1-P)}{(n_S+n)P^2(p_S - p'_S)^2} \% \quad (8)$$

where

$$G = \frac{n_T}{n_T+n} p'_S(1-\theta_S) + \frac{n}{n_T+n} (1-p'_S)\theta_S, \quad (9)$$

which lies between  $p'_S(1-\theta_S)$  and  $(1-p'_S)\theta_S$ .

Example 1. Using as 'typical' values of the probabilities  $p_S$ ,  $p'_S$  and  $P$  the values 0.9, 0.1 and 0.1 respectively we find that

$$G = (n_T+n)^{-1}(0.082n_T + 0.162n)$$

(so that  $G$  lies between 0.082 and 0.162) and the approximate proportional bias of  $\tilde{p}_T$  is between 0 and  $1406.25 G(n_S+n)^{-1}\%$ . Note that the upper limit is less than  $227.8 (n_S+n)^{-1}\%$ , so if  $n_S+n > 100$  the approximate proportional bias is less than 2.28%. The next section contains a numerical assessment of formula (5.2), without specifying values of  $p_T$  and  $p'_T$ .

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distrib	
Avail	
Dist	
A-1	



### 3. SOME NUMERICAL APPROXIMATIONS

Utilizing the reasonable assumption that  $p_S \gg p'_S$ , and neglecting terms in  $p'_S$  and  $p'^2_S$  in the numerator of (5.2) we find

$$\text{var}(\tilde{p}_T) \approx p_T^2 p^{-2} (p_S - p'_S)^{-2} \left\{ \frac{\phi(1-\phi)}{np_T^2} - \frac{2\phi(1-\theta_S)}{(n_S+n)p_T} + \frac{\theta_S(1-\theta_S)}{n_S+n} \right\} \quad (10)$$

Taking  $p_S = 0.9$ ,  $p'_S = 0.1$  so that  $\theta_S = 0.8P + 0.1$  and  $\phi = 0.9 p_T P + 0.1 p'_T(1-P)$ , we obtain from (10)

$$\begin{aligned} \text{var}(\tilde{p}_T) \approx & \frac{p_T^2}{0.64P^2} \left[ \frac{\{0.9p_T P + 0.1p'_T(1-P)\} \{1 - 0.9p_T P - 0.1p'_T(1-P)\}}{np_T^2} \right. \\ & \left. + \frac{(0.8P+0.1)(0.9-0.8P) - \{1.8p_T P + 0.2p'_T(1-P)\}(0.9-0.8P)p_T^{-1}}{n_S + n} \right] \end{aligned}$$

Now taking  $P = 0.1$ , we find

$$\text{var}(\tilde{p}_T) \approx \frac{p_T^2}{0.0064} \left[ \frac{0.09(p_T + p'_T) \{1 - 0.09(p_T + p'_T)\}}{np_T^2} - \frac{0.1476}{n_S + n} \cdot \frac{p'_T}{p_T} \right] \quad (11)$$

$$\approx \frac{14.375}{n} (p_T + p'_T) \{1 - 0.09(p_T + p'_T)\} - \frac{23.06}{n_S + n} \cdot \frac{p'_T}{p_T} \quad (12)$$

Since  $0.09(p_T + p'_T) \{1 - 0.09(p_T + p'_T)\} < \frac{1}{4}$  (because  $0.09(p_T + p'_T) < 1$ ) the right hand side of (12) is less than

$$(0.0256n)^{-1} < 39.1 n^{-1}.$$

In the next section we will compare the asymptotic variances and covariances of  $\tilde{\theta}_S, \tilde{\theta}_T$  and  $\tilde{P}$  with those for maximum likelihood estimators  $\hat{\theta}_S, \hat{\theta}_T$  and  $\hat{P}$  of  $\theta_S, \theta_T$  and  $P$  respectively.

#### 4. ANALYSIS II (Maximum Likelihood Estimators)

The likelihood function of  $Z_S$ ,  $Z_T$  and  $Z$  is

$$\begin{pmatrix} n_S \\ Z_S \end{pmatrix} \begin{pmatrix} n_T \\ Z_T \end{pmatrix} \begin{pmatrix} n \\ Z_{11}, Z_{10}, Z_{01}, Z_{00} \end{pmatrix} \theta_S^{Z_S} (1-\theta_S)^{n_S-Z_S} \theta_T^{Z_T} (1-\theta_T)^{n_T-Z_T} \\ \phi^{Z_{11}} (\theta_S - \phi)^{Z_{10}} (\theta_T - \phi)^{Z_{01}} (1-\theta_S - \theta_T + \phi)^{Z_{00}}$$

Equating derivatives of the log-likelihood to zero gives the following equations for  $\hat{\theta}_S$ ,  $\hat{\theta}_T$  and  $\hat{\phi}$ :

$$\frac{Z_S}{\hat{\theta}_S} - \frac{n_S - Z_S}{1 - \hat{\theta}_S} + \frac{Z_{10}}{\hat{\theta}_S - \hat{\phi}} - \frac{Z_{00}}{1 - \hat{\theta}_S - \hat{\theta}_T + \hat{\phi}} = 0 \quad (13.1)$$

$$\frac{Z_T}{\hat{\theta}_T} - \frac{n_T - Z_T}{1 - \hat{\theta}_T} + \frac{Z_{01}}{\hat{\theta}_T - \hat{\phi}} - \frac{Z_{00}}{1 - \hat{\theta}_S - \hat{\theta}_T + \hat{\phi}} = 0 \quad (13.2)$$

$$\frac{Z_{11}}{\hat{\phi}} - \frac{Z_{10}}{\hat{\theta}_S - \hat{\phi}} - \frac{Z_{01}}{\hat{\theta}_T - \hat{\phi}} + \frac{Z_{00}}{1 - \hat{\theta}_S - \hat{\theta}_T + \hat{\phi}} = 0 \quad (13.3)$$

subject to  $0 < \hat{\phi} < \hat{\theta}_S, \hat{\theta}_T < 1$  and  $\hat{\phi} > \hat{\theta}_S + \hat{\theta}_T - 1$ .

The information matrix is

$$V^{-1} = \begin{pmatrix} \frac{n_S}{\theta_S(1-\theta_S)} + \frac{n(1-\theta_T)}{(\theta_S - \phi)(1-\theta_S - \theta_T + \phi)} & \frac{n}{1-\theta_S - \theta_T + \phi} & -\frac{n(1-\theta_T)}{(\theta_S - \phi)(1-\theta_S - \theta_T + \phi)} \\ \frac{n}{1-\theta_S - \theta_T + \phi} & \frac{n_T}{\theta_T(1-\theta_T)} + \frac{n(1-\theta_S)}{(\theta_T - \phi)(1-\theta_S - \theta_T + \phi)} & -\frac{n(1-\theta_S)}{(\theta_T - \phi)(1-\theta_S - \theta_T + \phi)} \\ -\frac{n(1-\theta_T)}{(\theta_S - \phi)(1-\theta_S - \theta_T + \phi)} & -\frac{n(1-\theta_S)}{(\theta_T - \phi)(1-\theta_S - \theta_T + \phi)} & \frac{n\{\theta_S\theta_T(1-\theta_S - \theta_T) + 2\theta_S\theta_T\phi - \phi^2\}}{\phi(\theta_S - \phi)(\theta_T - \phi)(1-\theta_S - \theta_T + \phi)} \end{pmatrix} \quad (14)$$

The determinant is

$$|V^{-1}| = \frac{n}{\phi(\theta_S - \phi)(\theta_T - \phi)(1 - \theta_S - \theta_T + \phi)} \left[ \frac{n_S n_T \gamma}{\theta_S \theta_T (1 - \theta_S)(1 - \theta_T)} + n(n_S + n_T + n) \right]$$

with  $\gamma = \theta_S \theta_T (1 - \theta_S - \theta_T + \phi) - \phi(\phi - \theta_S \theta_T)$ .

And from the asymptotic variance-covariance matrix  $V$  we obtain

$$\begin{aligned} \text{var}(\hat{\theta}_S) &\doteq \frac{n}{|V^{-1}|} \left\{ \frac{n_T}{n_T(1-\theta_T)} \left[ \frac{1}{\phi} + \frac{1}{\theta_S - \phi} + \frac{1}{\theta_T - \phi} + \frac{1}{1 - \theta_S - \theta_T + \phi} \right] + n \left[ \frac{1}{\theta_T - \phi} + \frac{1}{1 - \theta_S - \theta_T + \phi} \right] \left[ \frac{1}{\phi} + \frac{1}{\theta_S - \phi} \right] \right\} \\ &\doteq \theta_S(1-\theta_S) (\gamma n_S^{-1} + \delta N^{-1}) / (\gamma + \delta) \end{aligned} \quad (15)$$

where  $N = n_S + n_T + n$  (= total number of observations) and

$$\delta = \frac{nN}{n_S n_T} \theta_S \theta_T (1 - \theta_S)(1 - \theta_T)$$

The MLE of  $P$  is

$$\hat{P} = (p_S - p'_S)^{-1} (\hat{\theta}_S - p'_S). \quad (16)$$

The asymptotic efficiency of  $\tilde{P}$  (see (4.1)) relative to  $\hat{P}$  is the same as that of  $\tilde{\theta}_S$  relative to  $\hat{\theta}_S$ , which is

$$100(n_S + n) (\gamma n_S^{-1} + \delta N^{-1}) / (\gamma + \delta) \% \quad (17)$$

Taking  $p_S = p_T = 0.9$ ,  $p'_S = p'_T = 0.1 = P$ , as in Example 1, and  $n_S = n_T = n$  ( $= \frac{1}{3}N$ ) we find  $\gamma = 0.0184680$  and  $\delta = 0.0653573$ , so (17) becomes

$$\begin{aligned} &100(n_S + n) (0.2203 n_S^{-1} + 0.7797 N^{-1}) \\ &= 2(0.2203 + 0.2599) = \underline{96.04\%}. \end{aligned}$$

The asymptotic variance of the MLE  $\hat{\phi}$  is

$$\text{var}(\hat{\phi}) \doteq \frac{1}{n} \frac{\delta \phi(1-\phi) - \delta \phi^2 N^{-1} \{n_S \theta_S^{-1}(1-\theta_S) + n_T \theta_T^{-1}(1-\theta_T)\} + \phi(\theta_S - \phi)(\theta_T - \phi)(1 - \theta_S - \theta_T + \phi)}{\delta + \gamma}$$

On the other hand, recalling that  $\text{var}(\tilde{\phi}) = n^{-1}\phi(1-\phi)$ , we find for the numerical values of the parameters used above, that the asymptotic efficiency of the moment estimator  $\tilde{\phi}$  is

$$100 \times \frac{0.0653573 \times 0.09\{0.91 - (2/3) \times 0.09 \times (0.18)^{-1} \times 0.82\} + 0.09 \times 0.09^2 \times 0.73}{0.09 \times 0.91(0.0653573 + 0.0184680)}$$

$$= 100 \times \frac{0.0037449 + 0.0005322}{0.0068653} = \underline{62.30\%}$$

The markedly lower asymptotic efficiency of  $\tilde{\phi}$  is associated with the fact that it does not utilize the information on values of  $\theta_S$  and  $\theta_T$  which is available from the other ( $n_S + n_T$ ) observations. Some support for this statement comes from the asymptotic efficiency of  $\tilde{\phi}$  if the values of  $\theta_S$  and  $\theta_T$  are known. This is

$$100 \times \frac{(\theta_S - \phi)(\theta_T - \phi)(1 - \theta_S - \theta_T + \phi)}{\gamma(1 - \phi)} \quad (19)$$

With the numerical values of  $\theta_S$ ,  $\theta_T$  and  $\phi$  which we have been using above this would give an asymptotic efficiency of only 35.18%.

##### 5. CALCULATION OF MAXIMUM LIKELIHOOD ESTIMATES

It is not possible to obtain explicit solutions of (13.1)-(13.3) for  $\theta_S, \theta_T$  and  $\phi$ , so a numerical solution must be sought.

An EM algorithm (see, e.g. Dempster et al. [2]) can be constructed in the following way. Introduce (unobserved) random variables  $Z_{ij}(S)$  ( $Z_{ij}(T)$ ) ( $i, j = 0, 1$ ) representing the numbers of  $i, j$  decision combinations which would have been obtained if the  $n_S$  ( $n_T$ ) individuals tested by  $S(T)$  had also been tested by  $T(S)$ . (Clearly

$$Z_{10}(S) + Z_{11}(S) = Z_S \quad \text{and} \quad Z_{01}(T) + Z_{11}(T) = Z_T.)$$

If values of these variables had been observed the maximum likelihood estimators would have been

$$\text{For } \theta_S: (Z_S + Z_{10(T)} + Z_{11(T)} + Z_{10} + Z_{11})N^{-1};$$

$$\text{For } \theta_T: (Z_{01(S)} + Z_{11(S)} + Z_T + Z_{01} + Z_{11})N^{-1};$$

$$\text{For } \phi: (Z_{11(S)} + Z_{11(T)} + Z_{11})N^{-1}.$$

Since

$$E[Z_{10(T)} | Z_T] = (n_T - Z_T)(\theta_S - \phi)(1 - \theta_T)^{-1}; \quad E[Z_{11(T)} | Z_T] = Z_T \phi \theta_T^{-1};$$

$$E[Z_{01(S)} | Z_S] = (n_S - Z_S)(\theta_T - \phi)(1 - \theta_S)^{-1}; \quad E[Z_{11(S)} | Z_S] = Z_S \phi \theta_S^{-1};$$

the EM algorithm leads to iteration from  $\theta_S^{(v)}$ ,  $\theta_T^{(v)}$ ,  $\phi^{(v)}$  to

$$\theta_S^{(v+1)} = N^{-1} \left[ Z_S + \frac{(n_T - Z_T)(\theta_S^{(v)} - \phi^{(v)})}{1 - \theta_T^{(v)}} + \frac{Z_T \phi^{(v)}}{\theta_T^{(v)}} + Z_{10} + Z_{11} \right] \quad (20.1)$$

$$\theta_T^{(v+1)} = N^{-1} \left[ \frac{(n_S - Z_S)(\theta_T^{(v)} - \phi^{(v)})}{1 - \theta_S^{(v)}} + \frac{Z_S \phi^{(v)}}{\theta_S^{(v)}} + Z_T + Z_{01} + Z_{11} \right] \quad (20.2)$$

$$\phi^{(v+1)} = N^{-1} \left[ \left[ \frac{Z_T}{\theta_T^{(v)}} + \frac{Z_S}{\theta_S^{(v)}} \right] \phi^{(v)} + Z_{11} \right] \quad (20.3)$$

Example 2. Table 1 sets out results of applying the EM algorithm to three illustrative sets of values of the n's and Z's. In each case  $n_S = n_T = n = 50$ ;

$Z_T = 10$ ;  $Z_{00} = 40$ ;  $Z_{01} = 3$ . The remaining values were

Set	$Z_S$	$Z_{10}$	$Z_{11}$
(I)	5	1	6
(II)	8	1	6
(III)	8	3	4

Table 1: EM algorithm solutions of equations (13.1)-(13.3)

Set $v$	(I)			(II)			(III)		
	$\theta_S^{(v)}$	$\theta_T^{(v)}$	$\phi^{(v)}$	$\theta_S^{(v)}$	$\theta_T^{(v)}$	$\phi^{(v)}$	$\theta_S^{(v)}$	$\theta_T^{(v)}$	$\phi^{(v)}$
0 (moment estimates)	0.1200	0.1900	0.1200	0.1500	0.1900	0.1200	0.1500	0.1700	0.0800
1	0.1221	0.1839	0.1154	0.1520	0.1924	0.1248	0.1539	0.1714	0.0865
2	0.1240	0.1816	0.1131	0.1522	0.1928	0.1270	0.1553	0.1714	0.0903
3	0.1251	0.1805	0.1119	0.1522	0.1929	0.1284	0.1560	0.1712	0.0928
4	0.1256	0.1800	0.1111	0.1522	0.1930	0.1294	0.1565	0.1711	0.0945
5	0.1259	0.1798	0.1106	0.1522	0.1930	0.1300	0.1568	0.1710	0.0957
6	0.1260	0.1797	0.1103	0.1522	0.1930	0.1305	0.1570	0.1709	0.0965
FINAL	0.1261	0.1796	0.1098	0.1523	0.1931	0.1315	0.1574	0.1707	0.0984

The initial values  $\theta_S^{(0)}$ ,  $\theta_T^{(0)}$  and  $\phi^{(0)}$  were the moment estimates. The table shows the results of the first six iterations and the final values, to four decimal places. (Speed of convergence can be improved, of course by using modified values of  $\theta_S^{(v)}$ ,  $\theta_T^{(v)}$ , and  $\phi^{(v)}$  for the  $(v+1)$ -th iteration, taking account of trends in values.)

The maximum likelihood estimates of  $P$ ,  $p_T$  and  $p_T'$  are obtained by replacing  $\theta_S$ ,  $\theta_T$  and  $\phi$  in (2.1)-(2.3) by their maximum likelihood estimates. We obtain the following formulas (provided the values lie between 0 and 1).

Set	$\hat{p}$	$\hat{p}_T$	$\hat{p}_T'$
(I)	$(0.1261 - p_S') / (p_S - p_S')$	$(0.1098 - 0.1796 p_S') / (0.1261 - p_S')$	$(0.1796 p_S - 0.1098) / (p_S - 0.1261)$
(II)	$(0.1523 - p_S') / (p_S - p_S')$	$(0.1315 - 0.1931 p_S') / (0.1523 - p_S')$	$(0.1931 p_S - 0.1315) / (p_S - 0.1523)$
(III)	$(0.1574 - p_S') / (p_S - p_S')$	$(0.0984 - 0.1707 p_S') / (0.1574 - p_S')$	$(0.1707 p_S' - 0.0984) / (p_S - 0.1574)$

In order to satisfy the conditions  $0 \leq \hat{P}, \hat{p}_T, \hat{p}'_T \leq 1$  we need

$p_S \geq \max(\hat{\theta}_S, \hat{\phi}/\hat{\theta}_T) \geq \min(\hat{\theta}_S, \hat{\phi}/\hat{\theta}_T) > p'_S$ . These conditions, for sets (I)-(III), are

$$(I) \quad p_S > 0.611; p'_S < 0.126$$

$$(II) \quad p_S > 0.681; p'_S < 0.152$$

$$(III) \quad p_S > 0.576; p'_S < 0.157.$$

[If the conditions are not met, then appropriate boundary values (0 if formula gives a negative value, 1 if it gives a value greater than 1) can be used.]

## 6. CONCLUDING REMARKS

The estimates of  $p_T, p'_T$  and  $P$  depend on the values assumed for  $p_S$  and  $p'_S$ . If these values are incorrect, biases will be introduced. The way in which the values used for  $p_S$  and  $p'_S$  affect the estimates can easily be appreciated from equations (2.1) - (2.3). For example, increase in either  $p_S$  or  $p'_S$  will tend to lead to negative bias in estimates of  $P$  (remembering that  $\theta_S < p_S$ ).

In this paper we have been concerned with estimation of  $p_T, p'_T$  (and also  $P$ ), supposing  $p_S, p'_S$  known. This has been effected via estimation of the parameters  $\theta_S, \theta_T$  and  $\phi$ . The same analysis can be used in other circumstances. For example, if  $P$  (proportion of nonconforming items) and  $p_S$  are known, then  $p_T, p'_T$  and  $p'_S$  can be estimated using the relationships

$$p'_S = (\theta_S - p_S P) (1-P)^{-1}$$

$$p_T = \frac{\phi(1-P) - (\theta_S - p_S P)\theta_T}{(p_S - \theta_S)P}$$

$$p'_T = \frac{p_S \theta_S - \phi}{p_S - \theta_S}$$

Of course, if  $P$  is known, as well as  $p_S$  and  $p'_S$ , then  $\theta_S$  is known and there is no need to take any observations with  $S$  alone - that is we can take  $n_S = 0$ .

#### Acknowledgement

Dr. Samuel Kotz's work was supported by the U.S. Office of Naval Research under Contract N00014-84-K-0301.

#### REFERENCES

- [1] Albers, W. and Veldman, H.J. "Adaptive Estimation of Binomial Probabilities under Misclassification," Statisti. Neerland., 38, 233-247. (1984).
- [2] Dempster, A.P., Laird, N.M. and Rubin, D.B. "Maximum Likelihood from Incomplete Data via the EM Algorithm," J.R. Statist. Soc., Ser. B, 39, 1-22. (1977)
- [3] Hochberg, Y. and Tenenbein, A. "On Triple Sampling Schemes for Estimating from Binomial Data with Misclassification Errors," Commun. Statist. - Theor. Math., 12, 1523-1533. (1983)
- [4] Johnson, N.L. and Kotz, S. Distributions in Statistics: Discrete Distributions, Wiley, New York. (1969)
- [5] Johnson, N.L. and Kotz, S. "Some Tests for Detection of Faulty Inspection," Statist. Hefte, 26, 19-29. (1985)
- [6] Tenenbein, A. "A Double Sampling Scheme for Estimating from Binomial Data with Misclassification," J. Amer. Statist. Assoc., 65, 1350-1362. (1970)



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

ADA170 320

## REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED			1b. RESTRICTIVE MARKINGS	
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; Distribution Unlimited	
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE				
4. PERFORMING ORGANIZATION REPORT NUMBER(S) UMI/MSS/1986/6			5. MONITORING ORGANIZATION REPORT NUMBER(S)	
6a. NAME OF PERFORMING ORGANIZATION Univ. of Maryland		6b. OFFICE SYMBOL (If applicable)		7a. NAME OF MONITORING ORGANIZATION Office of Naval Research
6c. ADDRESS (City, State, and ZIP Code) Dep't. of Management and Statistics University of Maryland College Park, Md. 20742			7b. ADDRESS (City, State, and ZIP Code)	
8a. NAME OF FUNDING / SPONSORING ORGANIZATION U.S. Office of Naval Research		8b. OFFICE SYMBOL (If applicable)		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-84-K-0301
8c. ADDRESS (City, State, and ZIP Code) Stat. and Probability Program Office of Naval Research Arlington, Va. 22217			10. SOURCE OF FUNDING NUMBERS	
			PROGRAM ELEMENT NO.	PROJECT NO.
			TASK NO.	WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification) Estimation from Binomial Data with Classifiers of Known and Unknown Imperfections				
12. PERSONAL AUTHOR(S) Norman L. Johnson and Samuel Kotz				
13a. TYPE OF REPORT UNCLASSIFIED		13b. TIME COVERED FROM 9/1/85 TO 9/1/86		14. DATE OF REPORT (Year, Month, Day) July 1986
15. PAGE COUNT 14				
16. SUPPLEMENTARY NOTATION				
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)	
FIELD	GROUP	SUB-GROUP		
			Binomial distribution; inspection errors; maximum likelihood; method of moments; EM algorithm	
19. ABSTRACT (Continue on reverse if necessary and identify by block number)				
<p>Observations from inspection by a "test" method and a standard method are combined to provide estimators of population proportion, and of probabilities of misclassification for the test method. Results are obtained in the case where the standard method is not perfect, but its misclassification probabilities have known values. Both moment and maximum likelihood estimators are considered and asymptotic properties of the resulting estimators are compared.</p>				
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED	
22a. NAME OF RESPONSIBLE INDIVIDUAL Samuel Kotz			22b. TELEPHONE (Include Area Code) (301) 454-6805	22c. OFFICE SYMBOL

END

DT/C

8-86